SOME RESULTS FOR IDENTIFICATION FOR SOURCES AND ITS EXTENSION TO LIAR MODELS

Zlatko Varbanov [1]
Department of Mathematics and Informatics,
Veliko Tarnovo University, 5000 Veliko Tarnovo,
(e-mail:vtgold@yahoo.com)

Introduction

The classical transmission problem deals with the question how many possible messages can we transmit over a noisy channel? Transmission means there is an answer to the question "What is the actual message?" In the identification problem we deal with the question how many possible messages the receiver of a noisy channel can identify? Identification means there is an answer to the question "Is the actual message u?". Here u can be any member of the set of possible messages.

Let $(\mathcal{U}, P)$ be a source, where $\mathcal{U} = \{1, 2, \ldots, N\}, P = \{P_1, P_2, \ldots, P_N\}$, and let $\mathcal{C} = \{c_1, c_2, \ldots, c_N\}$ be a binary prefix code (PC) for this source with $||c_u||$ as length of c_u. Introduce the random variable U with $\text{Prob}(U = u) = p_u$ for $u = 1, 2, \ldots, N$ and the random variable C with $C = c_u = (c_1, c_2, \ldots, c_{u||c_u||})$ if $U = u$.

We use the PC for noiseless identification, that is user u wants to know whether the source output equals u, that is, whether C equals c_u or not. The user iteratively checks whether C coincides with c_u in the first, second, etc. letter and stops when the first different letter occurs or when $C = c_u$.

What is the expected number $L_{\mathcal{C}}(P, u)$ of checkings?

In order to calculate this quantity we introduce for the binary tree $T_{\mathcal{C}}$, whose leaves are the codewords $c_1, c_2, \ldots, c_N$, the sets of leaves $\mathcal{C}_{ik}(1 \leq i \leq N; 1 \leq k)$, where $\mathcal{C}_{ik} = \{c \in \mathcal{C} : c$ coincides with c_i exactly until the k'th letter of $c_i\}$. If C takes a value in $\mathcal{C}_{uk}, 0 \leq k \leq ||c_u|| - 1$, the answers are k times "Yes" and 1 time "No". For $C = c_u$ the

$$L_{\mathcal{C}}(P, u) = \sum_{k=0}^{||c_u||-1} P(C \in \mathcal{C}_{uk})(k + 1) + ||c_u||P_u.$$

For a code $\mathcal{C}$

$$L_{\mathcal{C}}(P) = \max_{1 \geq u \geq N} L_{\mathcal{C}}(P, u)$$

is the expected number of checkings in the worst case and

$$L(P) = \min_{\mathcal{C}} L_{\mathcal{C}}(P)$$

is this number for the best code.

[1]Supported by COMBSTRU Research Training Network HPRN-CT-2002-00278

Dagstuhl Seminar Proceedings 06201
Combinatorial and Algorithmic Foundations of Pattern and Association Discovery
http://drops.dagstuhl.de/opus/volltexte/2006/782

Results for uniform distribution

Let $P^N = \{\frac{1}{N}, \ldots, \frac{1}{N}\}$. We construct a prefix code $\mathcal{C}$ in the following way. In each node (starting at the root) we split the number of remaining codewords in proportion as close as possible to $(\frac{1}{2}, \frac{1}{2})$.

It is known [1] that

$$\lim_{N\to\infty} L_{\mathcal{C}}(P^N) = 2 \tag{1}$$

Also, in [2] was stated the problem to estimate an universal constant $A = \sup L(P)$ for general $P = (P_1, \ldots, P_N)$. We compute this constant for uniform distribution and this code $\mathcal{C}$.

Using decomposition formula for subtrees, we obtain the following recursion

$$L_{\mathcal{C}}(P^N) = \frac{\lceil \frac{N}{2} \rceil}{N} L_{\mathcal{C}}(P^{\lceil \frac{N}{2} \rceil}) + 1, L_{\mathcal{C}}(P^2) = 1 \tag{2}$$

From (2) follows that the worst case for $L_{\mathcal{C}}(P^N)$ is when $N = 2^k + 1$, for any integer k. We compute the exact value for $L_{\mathcal{C}}(P^N)$ in this case and obtain

$$\sup_N L_{\mathcal{C}}(P^N) = 2 + \frac{log_2(N-1) - 2}{N}$$

Also, we consider the average number of checkings, if code $\mathcal{C}$ is used

$$L_{\mathcal{C}}(P, P) = \sum_{u\in\mathcal{U}} P_u L_{\mathcal{C}}(P, u) \text{ , and } L_{\mathcal{C}}(P^N, P^N) = \frac{1}{N} \sum_{u\in\mathcal{U}} L_{\mathcal{C}}(P^N, u)$$

is this number for uniform distribution.

We calculate the exact values of $L_{\mathcal{C}}(P^N)$ and $L_{\mathcal{C}}(P^N, P^N)$ for some N and summarize them in Table 1 (from [1] is known that for $N = 2^k, L_{\mathcal{C}}(P^N) = L_{\mathcal{C}}(P^N, P^N) = 2 - \frac{2}{N}$).

TABLE 1 - some exact values for uniform distribution, $2^k < N < 2^{k+1}, k \geq 3$

N	$L_{\mathcal{C}}(P^N)$	$L_{\mathcal{C}}(P^N, P^N)$
$2^k + 1$	$2 + \frac{log_2(N-1)-2}{N}$	$2 - \frac{2N - log_2(N-1)}{N^2}$
$2^k + 2^{k-1} - 1$	2	$2 - \frac{5(N+1) - 3log_2(\frac{2N+2}{3})}{3N^2}$
$2^k + 2^{k-1}$	$2 - \frac{1}{N}$	$2 - \frac{5}{3N}$
$2^k + 2^{k-1} + 1$	$2 + \frac{log_2(\frac{N-1}{12})}{N}$	$2 - \frac{(5N-2) - 3log_2(\frac{N-1}{12})}{3N^2}$
$2^{k+1} - 1$	$2 - \frac{1}{N}$	$2 - \frac{2N - log_2(N+1)+1}{N^2}$

Extension to liar models

Suppose that when user u iteratively checks whether C coincides with c_u in the first, second, etc. letter, for some reasons he obtains wrong information in any position. Then there is a lie(error) in this position of the codeword. In this model with lies, the user knows only that the general number of lies is at most e and no information for the positions of lies.

Let $L_{\mathcal{C}}(P,u) = L_{\mathcal{C}}(P)$ for any $u \in \mathcal{U}$. We denote by $L_{\mathcal{C}}(P;e)$ the expected number of checkings if there are at most e lies. In this case, the user needs of $e+1$ the same answers ("Yes" or "No") to be sure for the correct answer in any position. If the user u has made $2e+1$ questions for any position he gets exact information for the value in this position. Therefore, there exists trivial upper bound

$$L_{\mathcal{C}}(P;e) \leq (2e+1)L_{\mathcal{C}}(P)$$

Clearly, this upper bound can be improved by decreasing the number of remaining lies. The following algorithm can be used for any $u \in \mathcal{U}$:

Step 0: BEGIN $i := 1, Checkings := 0$, actual message $:= v$;
Step 1: If $i > ||c_v||$ then Step 3. Otherwise, check codeword position i until $e+1$ the same answers. Let t be the number of obtained answers "Yes" and f be the number of obtained answers "No";
Step 2: $Checkings := Checkings + (t+f)$. If $t > f$, then $e := e - f$, $i := i+1$, Step 1. Otherwise, the actual message $v \neq u$;
Step 3: END.

Let v be the current checked codeword and let i be the first position in which c_u and c_v differ (if $c_u = c_v$ then $i = ||c_u||$). We can see that the worst case with respect by e is when all lies(errors) occur in position i. In this case

$Checkings = (e+1)(i-1) + (2e+1).1 = e(i+1) + i.$

If there is a lie in any position m $(1 \leq m \leq i-1)$, for every position j $(m+1 \leq j \leq i)$ the user needs of e the same answers. Then

$Checkings = (m-1)(e+1) + (e+2) + (i-m-1)e + (2e-1) = e(i+1) + m < e(i+1) + i$

Therefore, if $k = ||c_u||$ and $P_{ui} = P(C \in \mathcal{C}_{ui})$, for the worst case we obtain the following upper bound

$$L_{\mathcal{C}}(P;e) \leq \sum_{i=0}^{k-1} P_{ui}(e(i+2)+i+1) + (e(k+1)+k)P_u = e\sum_{i=0}^{k-1} P_{ui}(i+2) + e(k+1)P_u + \sum_{i=0}^{k-1} P_{ui}(i+1) + kP_u$$

$$= e\sum_{i=0}^{k-1}(P_{ui}(i+1) + P_{ui}) + e(k+1)P_u + L_{\mathcal{C}}(P) = e(\sum_{i=0}^{k-1} P_{ui}(i+1) + kP_u) + e(\sum_{i=0}^{k-1} P_{ui} + P_u) + L_{\mathcal{C}}(P)$$

$$= eL_{\mathcal{C}}(P) + e + L_{\mathcal{C}}(P) = \underline{(e+1)L_{\mathcal{C}}(P) + e}$$

Let $M_{\mathcal{C}}(P;e) = (e+1)L_{\mathcal{C}}(P) + e$. Then from (1) follows that for uniform distribution P^N

$$\lim_{N\to\infty} M_{\mathcal{C}}(P^N;e) = 3e+2$$

Also, for general distribution $P = (P_1, P_2, \ldots, P_N)$ we know [1] that $L(P) \leq 3$. Therefore, for $L(P; e)$ (the expected number of checkings for the best code $\mathcal{C}$ and at most e lies) we obtain that

$$L(P; e) \leq 4e + 3$$

References

[1] R. Ahlswede, B. Balkenhol, and C. Kleinewächter, Identification for sources, preprint 00-120, SFB 343 "Diskrete Strukturen in der Mathematik", Bielefeld University, 2000

[2] R. Ahlswede, Identification entropy, preprint, 2004